

Learning for Visual Synthesis and Transformation

Xinyuan Chen

Faculty of Engineering and Information Technology

University of Technology Sydney

A thesis submitted for the degree of

Doctor of Philosophy

April 2020

I would like to dedicate this thesis to my loving parents
Lianfu Chen and *Meiyun Mao*

Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for the collaborative degree and/or fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This research is supported by the Australian Government Research Training Program.

This thesis is the result of a Collaborative Doctoral Research Degree program with Shanghai Jiao Tong University.

signature: Production Note:
Signature removed
prior to publication.

Acknowledgements

I would like to take this good opportunity to appreciate my advisors, several professors, my colleagues, my friends and my family for their significant help during my doctoral study in Shanghai Jiao Tong University.

My deepest gratitude goes first and foremost to my supervisor Prof. Dacheng Tao for his continuous support, unlimited patience and supportive guidance. He was always ready to help, and was very willing to teach everything he knows to me. His incredible enthusiasm and high demand motivate me to set up high academic research standards and submit papers to the leading journals or conferences in my research field.

I would also like to thank my supervisor Prof. Xiaokang Yang from Shanghai Jiao Tong University. He has given me not only plenty of freedom to explore in my research field, but also numerous constructive suggestions to help me out of various difficulties in study and life. I am particularly impressed by his encouraging attitude and expert knowledge for my research. Without his high academic standard, patient guidance, and unreserved support, I can never imagine I could have finished this thesis.

Besides my principal supervisors, I would like to give special thanks to my advisor: Dr. Chang Xu. I am grateful for his great patience in answering my simple and nearly endless questions from research motivation to implementation details, as well as improving my writing word-for-word. He taught me how to find interesting ideas, how to develop solid algorithms, and how to write technical papers all from scratch. I also thank Dr. Youming Qiao, Dr. Guoqiang Zhang

in UTS, Dr. Tongliang Liu, Dr. Shaoli Huang in USYD for their generous help during my study in Sydney.

I would also like to thank Prof. Li Song who led me into the field of computer vision. During my early days of academic research in Shanghai Jiao Tong University, he gave me valuable instructions and generous support. I also would like to thank Dr. Chao Ma. As a senior, he set up a great role for me. As an advisor, he provided me a lot of suggestions and help. Moreover, I would like to thank Prof. Bingbing Ni, Yi Xu and Guangtao Zhai for their advices and help.

The most enjoyable thing during my Ph.D. study is the opportunity to meet my dear colleagues and friends. I would like to thank Yilin Dong, Mengyue Shi, Yihang Huang, Han Zhang, Shaowei Xie, Zhe Ren, Genning Zhang for being close friends and offering me timely companionship; thank Muming Zhao, Dr. Jiangchao Yang, and Guo Lu for generous help and company in both Shanghai and Sydney; and thank Dr. Yichao Yan, Minsi Wang, Jingwei Xu, Wenbo Bao, Yunqian Wen, Jun Ling, Han Xue and many others.

Then comes my dear colleagues and friends in my doctoral study in Sydney: Dr. Erkun Yang, Dr. Chaoyue Wang, Yuxuan Du, Yali Du, Dr. Shan You, Dr. Jianfeng Dong, Yuangang Pan, Dr. Baosheng Yu, Dr. Liu Liu, Dr. Xiyu Yu, Dr. Huan Fu, Shanshan Zhao, Lianbo Zhang, Dalu Guo, Zeyu Feng, Zhe Chen, Rui Geng, Gengxing Wang and Xiaofei Liu. To my friends, thanks for all your support and company during my joyful and stressful time, and I cherish our friendship. To my colleague, it is fantastic to have the opportunity to work with you.

Finally, I want to dedicate this thesis to my parents who gave me unreserved love and support. Every Friday night, you always wait for my phone to hear my experience of study and life, share my joys and sorrows. Because of you, I never feel alone and frustrated through my difficult times. Thanks for always being there for me.

Abstract

Visual synthesis is one of the most fundamental problems in computer vision and artificial intelligence. Visual synthesis aims to create pixel-level data (e.g., images and videos) based on descriptions such as texts, noise, semantic annotations and images. Recently, deep generative learning has greatly promoted the development of visual synthesis. However, the existing generative methods still suffer from several issues, including model interpretation, controllability, stability, efficiency and performance. In this thesis, several generative models are proposed to address these challenges. This thesis makes the following contributions:

First, this thesis introduces an attention generative model for local image synthesis, so as to improve the controllability and interpretation of the generative model. How to precisely locate the foreground region in the image and generate the target object to the specified region is the key problem in the local image synthesis task. The object transfiguration task is an application of the local image synthesis, which aims to transform the object of images to another object. Existing generative methods often fail to decompose the foreground and background. In this thesis, the attention mechanism is incorporated into generative models, so as to transform the object of our interests without altering the background. The model is built by decomposing the generative network into two separate networks, each of which is dedicated to one sub-task: to detect the region of interests and to generate the object from one object to another. The attention network predicts spatial attention maps of images, and the transformation network focuses on translating objects. The attention network produces attention maps which are encouraged to be sparse so that the model

can only pay attention to the objects of interest. Also, a novel perceptual loss is introduced to improve the quality of transformed images in the high-level feature space. Experimental results demonstrate the necessity of investigating attention in image-to-image transformation, and that the improvement of the quality of generated images.

Second, this thesis proposes a multi-domain generative model that multiple styles of images can be generated in a single network. The major challenge is how to efficiently generate multiple styles in a single network. Our model is capable to extract the content and style feature of images, and apply multiple style features to the content image. This thesis proposes a gated generative model that consists of three modules: an encoder, a gated transformer, and a decoder. Different styles can be achieved through different branches of gated transformers while the encoder and decoder are used for capturing structure information sharing weights for all styles. A discriminative network is used to distinguish whether the input image is a stylized or genuine image. An auxiliary classifier is used to recognize the style categories of transferred images, thereby helping to generate images in multiple styles. In addition, to stabilize the adversarial training process, an auto-encoder reconstruction loss is introduced by combining the encoder and decoder module. Extensive experiments demonstrate the stability and effectiveness of the proposed model for multi-domain image synthesis.

Third, this thesis investigates the video synthesis problem on the long-term horizon. A temporal generative model is proposed for long-term video frame prediction. The existing generative model for video prediction usually cannot output high-quality predictions for a long-time horizon. The reason is that those methods recursively output subsequent frames by taking the newly generated frames as observations, consequently the prediction error accumulates dramatically. The introduced retrospection process is designed to look back on what has been learned from the past and rectify the prediction deficiencies.

To this end, a retrospection network is built to reconstruct the past frames given the currently predicted frames. On the other hand, an auxiliary route is built by reversing the flow of time and executing a similar retrospection. These two routes interact with each other to boost the performance of retrospection network and enhance the understanding of dynamics across frames, especially for the long-term horizon.

Overall, this thesis investigates the deep generative model and solves several practical issues for visual synthesis and transformation. For local image synthesis, we propose an attention generative model. We also propose a gated generative model for generating multi-domain of images in a single generative network. For video synthesis, a temporal generative model is proposed to output long-term video frames by incorporating the prediction and retrospection process in the model. Extensive experimental results on large-scale benchmark datasets demonstrate that the proposed methods in this thesis perform favorably against previous visual synthesis algorithms in terms of efficiency, controllability, and robustness.

Contents

Contents	xi
List of Figures	xiv
1 Introduction	1
1.1 Background	1
1.2 Overview	3
1.2.1 Deep Generative Models	4
1.2.2 Visual Synthesis and Transformation	5
1.2.3 Challenges	8
1.3 Contributions	9
1.4 Organization	10
2 Literature Review	13
2.1 Generative Models	13
2.2 Generative Adversarial Networks	14
2.3 Image Synthesis	16
2.4 Video Synthesis	18
3 Attention Generative Model for Local Image Synthesis	20
3.1 Introduction	21
3.2 Related Work	25
3.2.1 Generative Adversarial Networks	25
3.2.2 Image-to-Image Transformation	26
3.2.3 Attention Model in Networks	27

CONTENTS

3.3	Preliminaries	28
3.4	Attention Generative Model	29
3.4.1	Generative Model	30
3.4.2	Attention Loss	32
3.4.3	Perceptual loss	33
3.4.4	Extra Supervision	34
3.5	Implementation	35
3.6	Experiments	37
3.6.1	Qualitative Comparisons	40
3.6.2	Background Consistency Comparison	41
3.6.3	Human Perceptual Study	44
3.6.4	Model Analysis	44
3.6.5	Comparison of Supervised Results	48
3.6.6	Global Image Transformation	50
3.7	Summary	50
4	Gated Generative Model for Global Image Transformation	52
4.1	Introduction	53
4.2	Related Work	56
4.2.1	Traditional Texture Transfer Method	56
4.2.2	Optimization-based Methods	57
4.2.3	Feedforward Networks-based Methods	57
4.2.4	Adversarial Network-based Methods	58
4.3	Gated Generative Model	59
4.3.1	Adversarial Network for Style Transfer	61
4.3.2	Auto-encoder Reconstruction Loss for Training Stabilization	61
4.3.3	Adversarial Gated Network for Multi-Style Transfer	62
4.4	Implementation	64
4.4.1	Network Configuration	64
4.4.2	Training Strategy	65
4.5	Experiments	67
4.5.1	Assessment of Image Quality	67
4.5.2	Texture synthesis	67

CONTENTS

4.5.3	Style Transfer	69
4.5.4	Analysis of Loss Function	76
4.5.5	Analysis of network architecture	79
4.5.6	Incremental Training	81
4.5.7	Linear Interpolation of Styles	82
4.6	Conclusions	83
5	Temporal Generative Model for Long-term Video Frame Synthesis	84
5.1	Introduction	85
5.2	Related Work	87
5.3	Temporal Generative Model	89
5.3.1	Preliminaries: Prediction Process	91
5.3.2	Retrospection Process	92
5.3.3	Full Objective	95
5.4	Implementation	96
5.4.1	Network Configuration	96
5.4.2	Training Strategy	99
5.5	Experiments	99
5.5.1	KTH and Weizmann Action Datasets	101
5.5.2	UCF-101 Dataset	105
5.5.3	Model Analysis	107
5.6	Retrospection for Other Models	111
5.7	Conclusion	112
6	Conclusions	114
6.1	Summary of Conclusions	114
6.2	Future Works	115
	References	117
	Publications	140

List of Figures

1.1	Computer vision and deep learning algorithms have shown great progress in visual understanding, e.g., image caption [147]. Our work focuses on an opposite way, visual synthesis with the goal of generating pixel-level visual data from abstraction concept.	2
1.2	The applications of image synthesis given input of image. Local image synthesis refers to generate or manipulate certain parts of images while keeping the remaining region consistent. Global image synthesis refers to generate images as a whole.	6
1.3	The organization overview of this thesis.	11
2.1	The illustration of the adversarial training procedure.	15
3.1	Comparisons of object transfiguration examples. From left to right: the input images, the transformed results of the prior model, and the transformed results of our proposed model. a) An example of horse \rightarrow zebra; b) An example of zebra \rightarrow horse.	21
3.2	Results of object transfiguration on different tasks: horse \leftrightarrow zebra, leopard \leftrightarrow tiger and apple \leftrightarrow orange. In each case, the first image is the original image, the second image is the synthesized image, and the third image is the predicted attention map. Our proposed model only manipulates the attention parts of the image and preserves the background consistency.	22

LIST OF FIGURES

3.3	The architecture of the proposed method. For clarity the target cycle is omitted. The grey dotted frame represents the generator of our model, which consists of a transformation network and an attention network. Detailed illustration of the generator is shown in Figure 3.4. The perceptual loss minimizes the distance in feature space between the transformed image and the overall images of the target domain.	30
3.4	The generator of Attention-GAN transforms the source image from one class to another. The attention network predicts the attention maps. The transformation network synthesizes the target object. A layered operation is applied to the background and transformed images to output the resulting image.	31
3.5	Comparison with CycleGAN [183], UAIT [101] and Attention-GAN [24] on horse \leftrightarrow zebra. In each case, the first image is the input image, the second is the result of CycleGAN, the third is the results of UAIT, the forth is the result of Attention-GAN and the last is the result of our Attention-GAN+.	39
3.6	Comparison with CycleGAN on apple \leftrightarrow orange and tiger \leftrightarrow leopard. In each case: input image (left), result of CycleGAN [183] (middle), and result of our Attention-GAN (right).	43
3.7	The stacked bar chart of participants' preferences for our methods compared to CycleGAN [173]. The blue bar indicates the number of images that more participants prefer our results. The gray bar indicates the number of images that more participants prefer CycleGAN's results. The orange bar indicates the number of images where two methods get an equal number of votes from 10 participants.	43
3.8	Generation results of our model on horse \rightarrow zebra. From left to right: Inputs, attention maps, outputs of transformation network, background images factorized by attention maps, object of images factorized by attention maps, final composite images.	45

LIST OF FIGURES

3.9	The effect of sparse loss with different parameters λ_{sparse} for mapping horse \rightarrow zebra. From left to right: input, output and attention map without sparse loss, input and attention map when $\lambda_{sparse} = 1$, input and attention map when $\lambda_{sparse}=5$	46
3.10	The results of tiger \leftrightarrow leopard with different values on $\lambda_{sparse} = \{0.1, 0.3, 0.5, 1\}$	46
3.11	Results of horse \leftrightarrow zebra by unsupervised Attention-GANs. From left to right: input images, outputs of the proposed Attention-GAN, the predicted attention maps of Attention-GAN.	47
3.12	Results of tiger \leftrightarrow leopard by unsupervised Attention-GANs. From left to right: input images, outputs of the proposed Attention-GAN, the predicted attention maps of Attention-GAN.	48
3.13	Results of apple \leftrightarrow orange by unsupervised Attention-GANs. From left to right: input images, outputs of the proposed Attention-GAN, the predicted attention maps of Attention-GAN.	49
3.14	Comparison of horse \rightarrow zebra between CycleGAN [183], unsupervised Attention-GAN, supervised Attention-GAN, and supervised Attention-GAN+.	50
3.15	Results of Summer \rightarrow Winter comparing with CycleGAN. From left to right: input images, results of CycleGAN, final outputs of the proposed Attention-GAN, the predicted attention maps of Attention-GAN.	51
4.1	Gated-GAN for multi-collection style transfer. The images are produced from a single model with a shared encoder and decoder are shared. Styles are controlled by switching different gated-transformer module. From left to right: original images, transferred images in Monet style, transferred images in Van Gogh's style, transferred images in Cezanne's style, transferred images in Ukiyoe-e's style.	54

LIST OF FIGURES

4.2	The architecture of the proposed adversarial gated networks: a generative network and a discriminative network. The generative network consists of three modules: an encoder, a gated transformer, and a decoder. Images are generated to different styles through branches in the gated transformer module. The discriminative network uses the adversarial loss to distinguish between stylized and real images. An auxiliary classifier supervises the discriminative network to classify the style categories.	60
4.3	Four cases of texture synthesis using Gated-GAN. For each case, the first column shows examples of texture, and the other three are synthesized results given different samples of Gaussian noise as inputs.	68
4.4	Visualization of learned features in the gated transformer of the generative networks. In each case, the left shows synthesized images and the right shows the corresponding features.	68
4.5	Collection style transfer on Photo \rightarrow Monet. From left to right: input photos, Monet's paintings picked from a similar landscape theme, and our stylized images. The photo is transferred adaptively based on different themes.	69
4.6	A four-style transfer network is trained to capture the styles of Monet, Van Gogh, Cezanne, and Ukiyo-e.	70
4.7	Comparison of our methods with image style transfer [40] on photo \rightarrow Monet and photo \rightarrow Ukiyo-e. From left to right: input photos, Gatys <i>et al.</i> 's results using different target style images, Gatys <i>et al.</i> 's results using the entire collection of artist and genre, our results for collection style transfer.	72
4.8	Comparison of our methods with universal style transfer [81] on photo \rightarrow Monet. From left to right: input images, results of [81] with the style image: Monet <i>Charing Cross Bridge</i> , results of [81] with the style image: Monet <i>Flowers at Vetheuil</i> , and our results of Monet's collection style transfer.	73

LIST OF FIGURES

4.9	Comparison with CycleGAN [183]. From left to right: original images, stylized images in Monet’s style, stylized images in Van Gogh’s style, stylized images in Cezanne’s style, stylized images in Ukiyo-e style. In each case, the first row shows the results produced by CycleGAN, and the second row shows our results.	74
4.10	Model size. We compare the number of parameters between our model and CycleGAN [183]. The x-axis indicates style number and the y-axis indicates the model size.	75
4.11	Comparison of our methods with Condition GAN and its variant. From left to right: input, condition GAN and condition GAN + cycle-consistent loss. Each row indicates different styles, from top to bottom: Monet, Ukiyo-e, Cezanne.	76
4.12	Qualitative comparison of the influence of parameter λ_{CLS} . The first column shows the input images. The rest columns demonstrate results with $\lambda_{CLS} = \{0, 0.1, 1, 10\}$. Each row demonstrates images transferred by different styles. From top to bottom: Monet, Cezanne, Van Gogh.	77
4.13	Qualitative comparison of the influence of parameter λ_R . The first column shows the input images. The rest columns demonstrate results with $\lambda_R = \{1, 5, 10, 20\}$. Each row demonstrates images transferred by different styles. From top to bottom: Monet, Cezanne, Ukiyo-e.	78
4.14	Comparison with a variant of our method across different training iterations for mapping images to Cezanne’s style. From left to right: original images, results after training for 10k, 100k, and 300k iterations with and without auto-encoder reconstruction loss.	80
4.15	Qualitative comparison of the influence of different network structures. The first row is the results of photo \rightarrow Cezanne, and the second row is the results of photo \rightarrow Van Gogh.	81
4.16	Comparison of incremental training. From left to right: original inputs, results of CycleGAN [14], results of our methods that all the styles are trained simultaneously, results of incremental training.	82

LIST OF FIGURES

4.17	Style interpolation. The leftmost image is generated in Monet’s style, and the rightmost image is generated in Van Gogh’s style. Images in the middle are convex combinations of the two styles. .	83
5.1	The illustration of our model. The generative model predicts the next frame conditioned on the previous frames (blue lines), while our model introduces a retrospection process to reconstruct the original input frames given predictions in a reverse chronological order (green lines).	91
5.2	The overall architecture of the proposed network. Left bottom: our model contains two routes: one is $G \rightarrow F$, and the other is $F \rightarrow G$. Right top: illustration of the route $G \rightarrow F$ in detail. In each route, our model consists of two processes. The prediction process executes recursively by taking the observations to generate subsequent frames, while the retrospection process synthesizes frames by observing the predicted frames in a backward manner. The discriminative network uses an adversarial loss to distinguish between predicted and real frames.	93
5.3	Qualitative comparison between our methods, MCNET and SVG on the KTH dataset. The top case corresponds to the action of boxing, and the lower case corresponds to the action of hand-waving.	102
5.4	Quantitative comparisons between different variants of our method and MCNET baseline in terms of PSNR, SSIM and LPIPS on the KTH dataset. “Ours” denotes our method (MCNET+Retrospection) with full objective. “Route F to G ” represents Route $F \rightarrow G$ alone (Equation 5.19). “Route G to F ” indicates Route $G \rightarrow F$ alone (Equation 5.20). Given 10 input frames, the models predict 100 frames recursively. For PSNR and SSIM, higher is better. For LPIPS, lower is better.	103
5.5	Quantitative comparisons between our method and MCNET in terms of PSNR, SSIM on the Weizmann dataset.	103

LIST OF FIGURES

5.6	Comparison with MCNET [144] in terms of the recognition rate. The recognition rate of the person detector that a person is recognized in the predicted frame.	104
5.7	Quantitative comparison on the UCF-101 dataset. MCNET [144] trained to predict 10 frames is denoted as “MCNET-T10”. The results are predicted by observing 10 previous frames. Our method less artifact and blur around the ambiguity region. The remarkable region is denoted in color and scaled.	105
5.8	Quantitative comparisons between our model, MCNET [144] and MCNET trained by 10 input frame and 10 output frames (indicated by “MCNET-T10”). In the test phase, the models predict 100 frames recursively given 10 input frames.	106
5.9	The stacked bar chart of participants preferences for our methods compared to MCNET [144]. The blue bar indicates the number of videos that more participants prefer our results. The gray bar indicates the number of videos that more participants prefer MCNET’s results. The orange bar indicates the number of videos where two methods get a equal number of votes.	107
5.10	Quantitative comparison of the retrospection loss with different parameter γ	109
5.11	Qualitative analysis for the function of the adversarial loss. First Row: ground-truth frames; Second Row: our results with full objective; Third Row: results without adversarial loss in Equation 5.21.	110
5.12	Ablation study for the function of the adversarial loss in terms of the recognition rate. The recognition rate of the person detector that a person is recognized in the predicted frame.	110
5.13	Quantitative comparisons between our method, SAVG, SVG w/o the retrospection process in terms of PSNR, SSIM and LPIPS. . .	111
5.14	Qualitative comparison between SVG and “SVG+retro” on the action of hand-waving. “SVG+retro” incorporating the retrospection process.	112